Remote Sensing Data Mining: New Problems and Research Opportunities

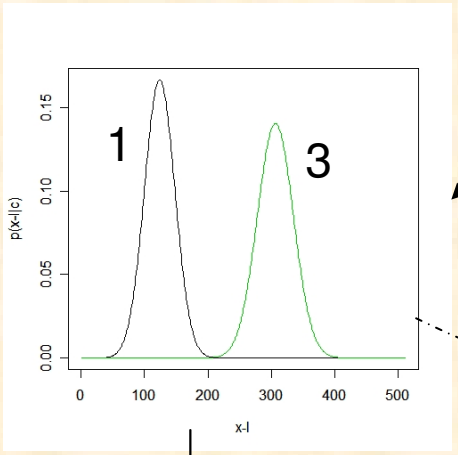
Ranga Raju Vatsavai (vatsavairr@ornl.gov)
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Collaborators:
Shashi Shekhar (CS/UMN)
Thomas E. Burk (RSL/UMN)
Budhendra Bhaduri (GIST/ORNL)

RS Data Mining

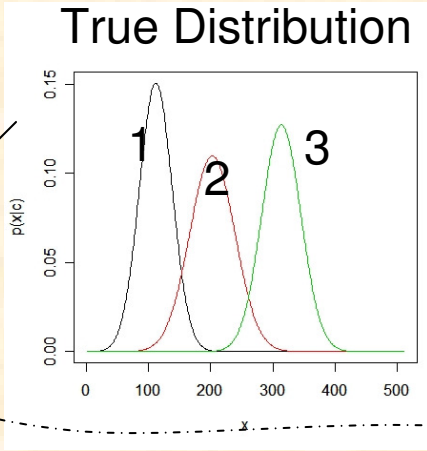
- **Increasing spectral, spatial, and temporal resolutions**
- **Challenges**
  - **Insufficient no of ground truth training samples**
  - **Aggregate Classes**
  - **Overlapping Classes**
- **New Approaches**
  - **Semi-supervised Learning**
  - **Sub-class classification**
  - **Mixture models for multi-source data**

Semi-supervised Learning

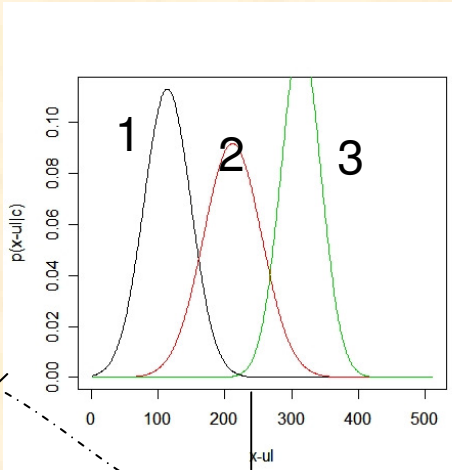
Labeled



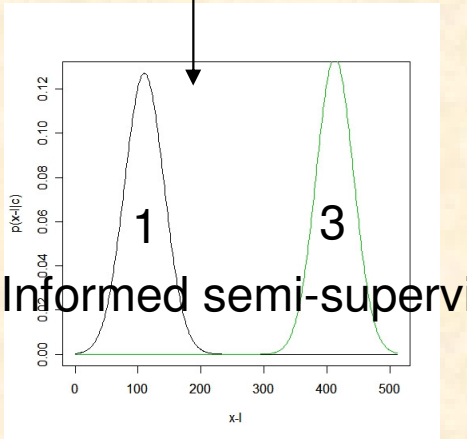
True Distribution



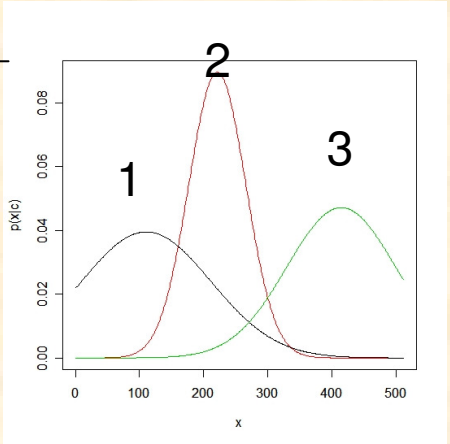
Unlabeled



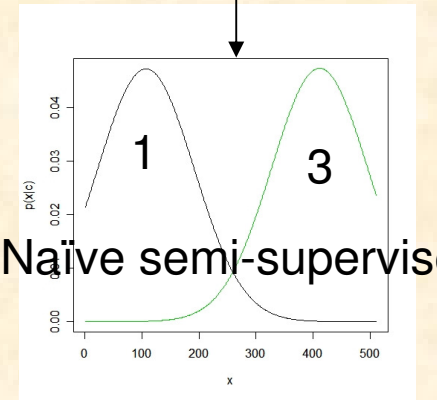
Match



Informed semi-supervised

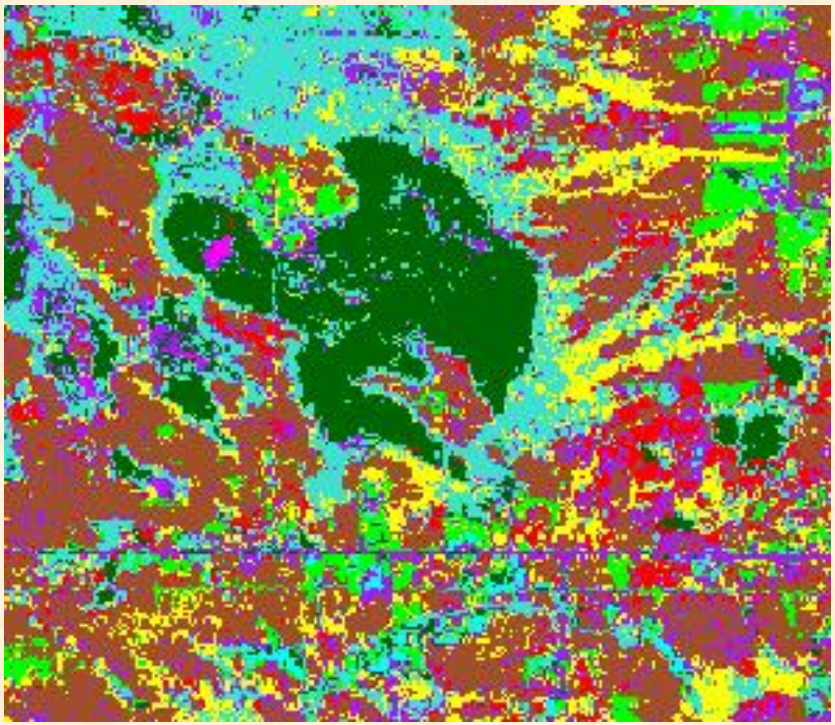


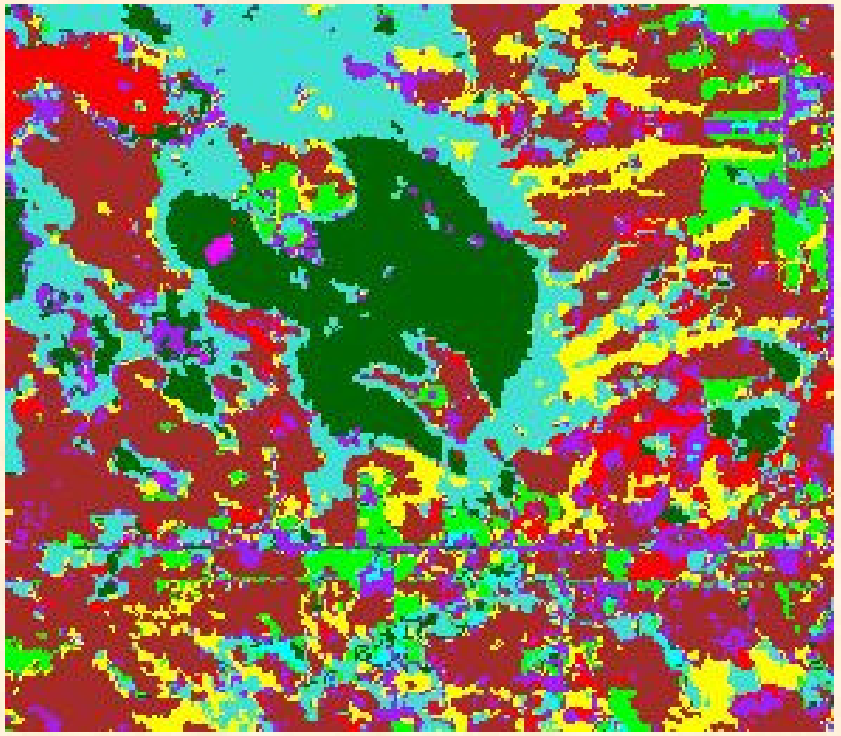
Clustering



Naïve semi-supervised

Spatial Semi-supervised Learning





Spatial Autocorrelation is very important, ignoring may result in salt and pepper noise

Sub-classes and Overlapping Classes

- Training is domain specific (Same image but different classes in different applications)
  - Forester: Hardwood, Conifer, …, Rest of image (few broad classes (Agriculture, Urban, Water)
  - Agriculture: Soybean, Wheat, … Rest (forest, Urban, …)
    - Violates basic assumption of unimodal Gaussian/class
- Mismatch between thematic classes and image classes
  - Upland hardwood vs. Low-land hardwood
  - High-density Urban vs. Low-density Urban

Sub-class Classification from Aggregate Class Labels

Each Class is Unimodal Gaussian

MLE

Each Aggregate Class is GMM

How many components?

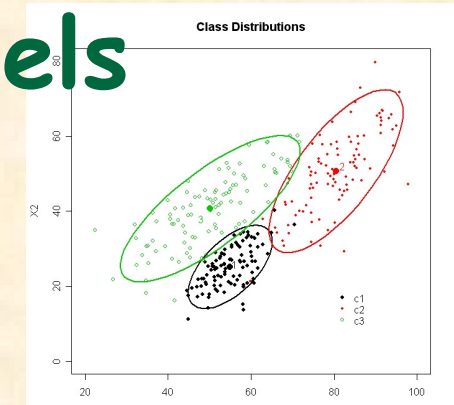
BIC/AIC Model Selection
+ Parameter Estimation

What are these components?

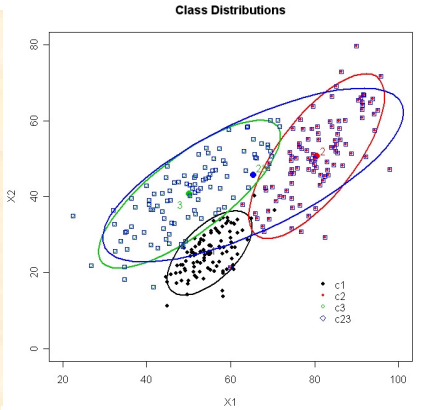
Few labels/
sub-class

Unlabeled
Samples

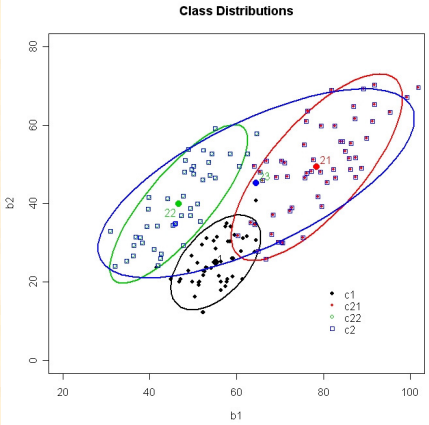
Semi-supervised Sub-class Classification



Actual Distribution of Classes



User given (Aggregate Class)



Sub-classes Recovered from Agg. Cl.

Multi-source Classification

- **Mixture Model**

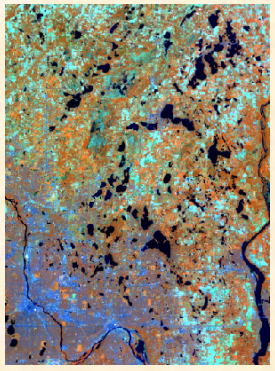
$$p(x \mid \theta) = \sum_{i=1}^{M} \alpha_i p_i(x \mid \theta_i) \quad [p_i(.) - Gaussian]$$

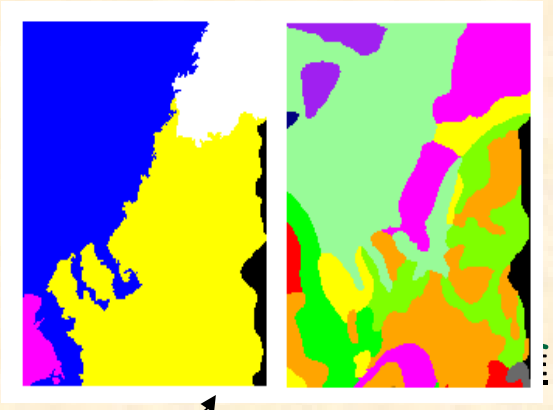
- **Mixture Model for Multi-source Data**

$$p(x_i \mid \theta) = \sum_{j=1}^{M} \alpha_j \prod_{l=1}^{2} p_{jl}(x_{il} \mid \theta_{jl})$$

$l = 1$: Continuous

$l = 2$: Discrete





Conclusions and Future Directions

- **Semi-supervised**
  - **Improved accuracy for as few as 2 labeled samples per class, Spatial classification is better**
  - **Challenge – Significance Testing + Matching**
  - **Future needs – "pixels" to "objects"**
- **Sub-class Classification**
  - **Can be used to recognize large number of finer classes**
  - **Challenge – collecting (few) labels for sub-classes**
- **Mixture Model for Multi-source Data**
  - **Very flexible framework**
  - **Challenge – variable selection, stratification, …**
- **We are working on**
  - **Image characterization (extending natural scene statistics)**
  - **Object-based classification, semantic labeling, image retrieval, …**